# Differential Privacy Webinar: What GIS users need to know

**Get Answers:** Questions will be answered at the end.

You can submit them at anytime through the Q&A module.

**On-Demand:** The recording will be posted shortly after the webinar.

You'll receive an email with the link to view or download.

**Start time:** The webinar will start at 8:00 am PDT

**Contact us:** For anything else, please email us:

**lpeters@esri.com**

# Differential Privacy:

## What GIS users need to know

## Today's Presenters

**Dr. John Abowd**
**Associate Director for Research & Methodology and Chief Scientist | US Census Bureau**

**Dr. Lauren Scott Griffin**
**Product Engineer and GIS Analyst**
**Spatial Statistics | Esri**

## Today's Moderator

**Linda Peters**
**Global Business Development**
**Official Statistics | Esri**

# Differential Privacy Webinar: What GIS users need to know

- What is Differential Privacy and Disclosure Avoidance
- How Differential Privacy might impact your work
- Dr. John Abowd:  US Census Bureau’s Disclosure Avoidance
- Discussion with Dr. Lauren Scott Griffin on key issues
- Open Q/A

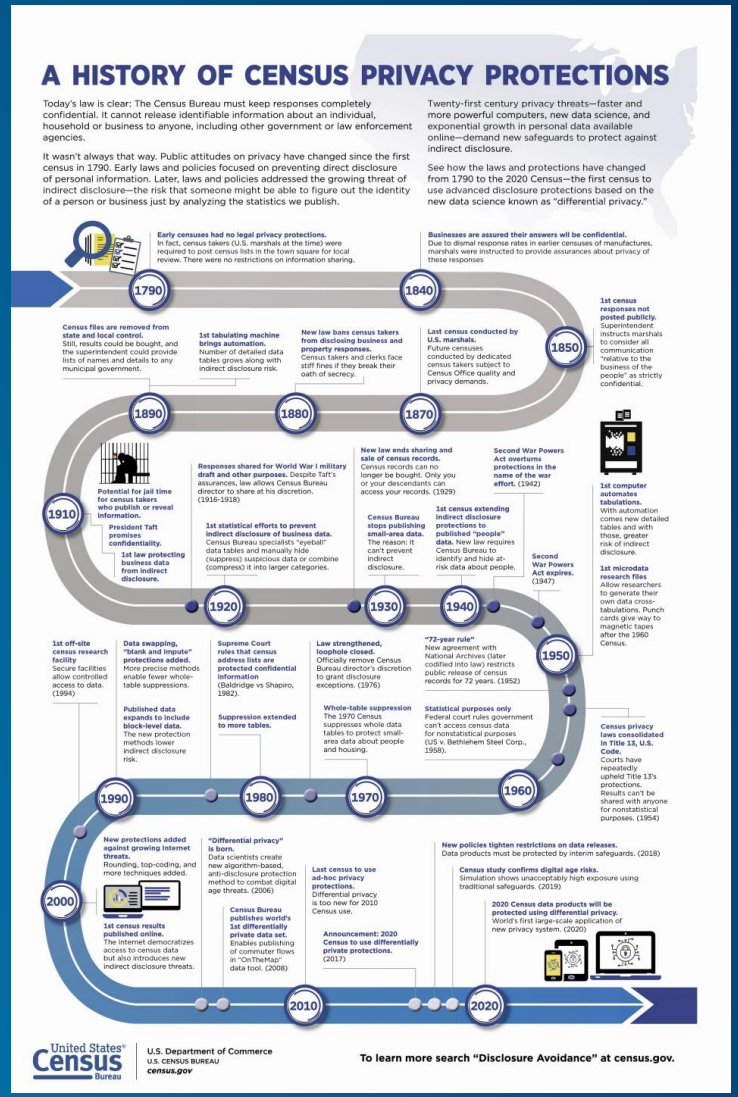
# Why Differential Privacy and Disclosure Avoidance

- **Census Bureau requirements**
  - **Constitutional mandate to enumerate the US population every ten years**
  - **Title 13: Protect individuals from being identified in published data.**
    - **These requirements are complex on their own and when viewed together are *at odds with one another***
- **Disclosure Avoidance**
  - **Prior censuses used various forms of disclosure avoidance to ensure that privacy is protected while also releasing high quality data that fits the needs for numerous use cases.**
    - ***Data Swapping and table suppression***

# What is Differential Privacy

- Differential privacy is a “formal privacy” approach that provides proven mathematical privacy assurances by adding uncertainty or “noise” to the released data.
- With differential privacy the “acceptable risk” can be quantified through a measure called Epsilon. When Epsilon is set to zero, the data are completely scrambled. When Epsilon is set to infinity, the data are as enumerated

***The Census Bureau will determine the amount of noise necessary to balance privacy loss and accuracy for every table released***

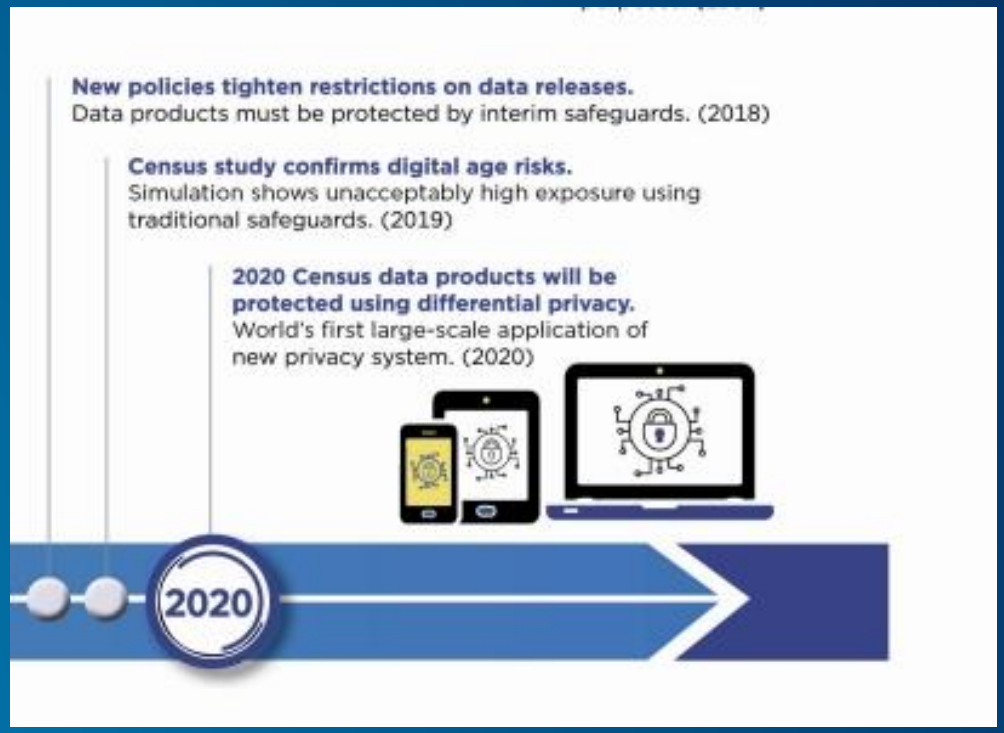


# What is Disclosure Avoidance System?

- If it were okay to have negative values and fractions (the result of injecting the DP noise), all would be fine.

- Because negative values and fractions are not acceptable (and because the smaller geographies need to add up to the larger geographies), the Census Bureau uses optimization methods to fix this in post processing.

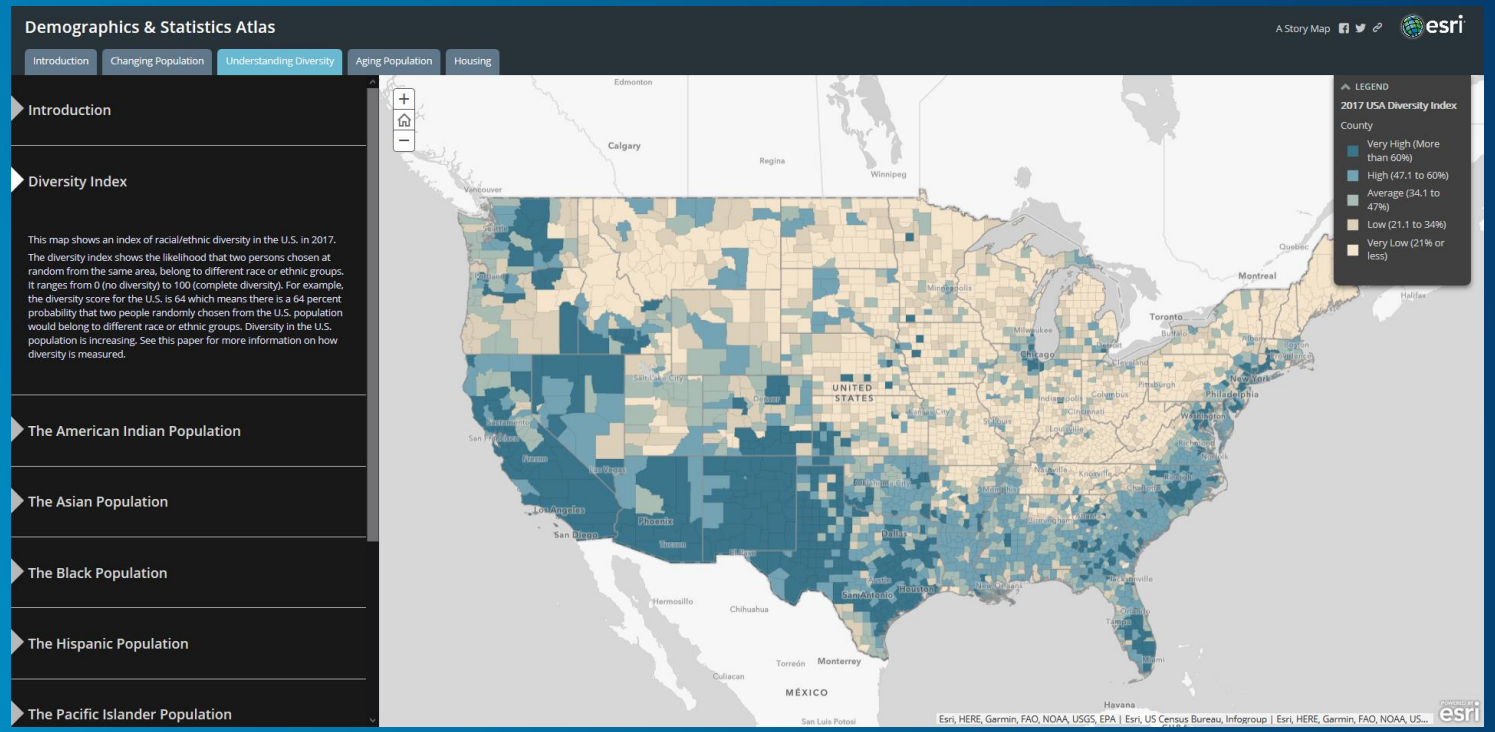
- Together, differential privacy and post-processing alterations are referred to as the 2020 Disclosure Avoidance System (DAS).



# How Differential Privacy might impact your work

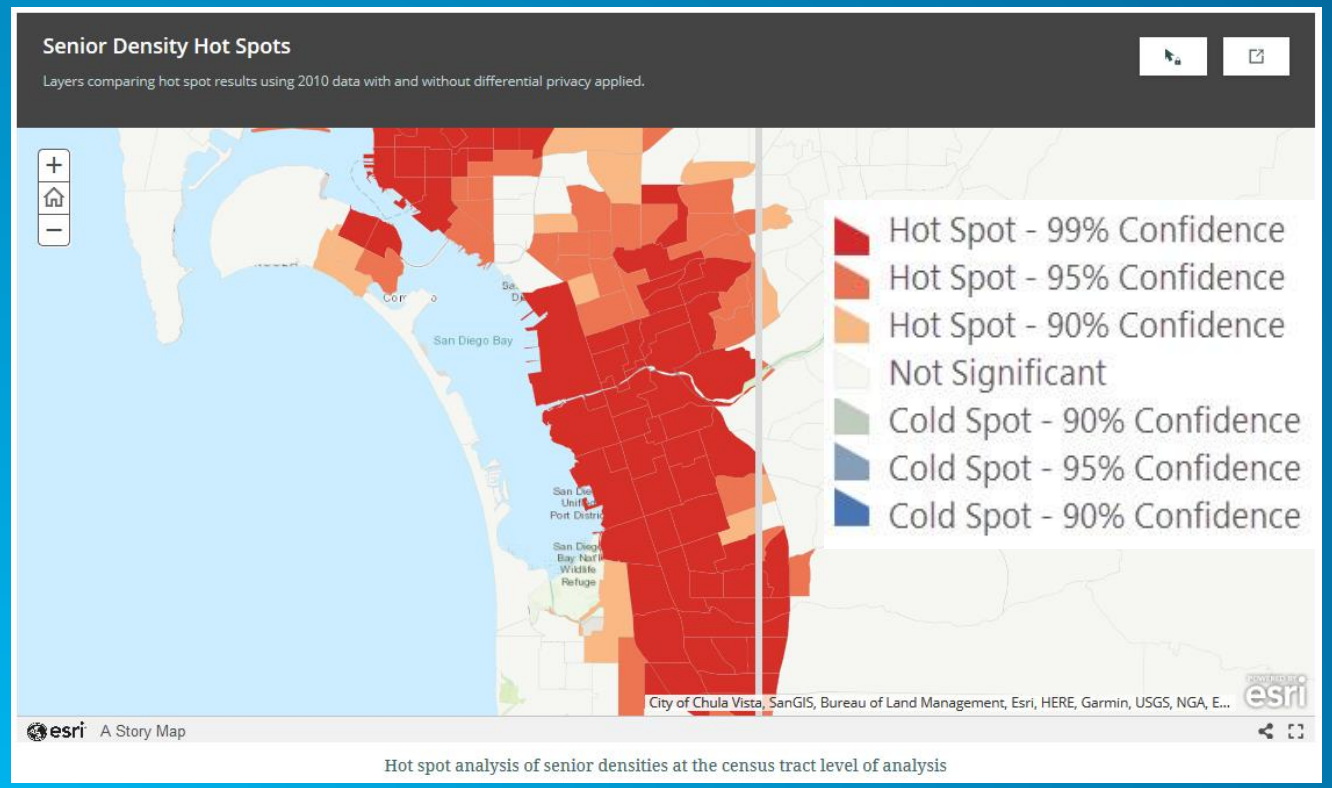
***Question: At what level of geography will summary files be released?***

- Accuracy
- Impact on Small Area (sub county) analysis
- Tracking marginalized populations
- Mapping Spatial Patterns
- Temporal Analysis
- Transparency
- Beyond 2020 what's next?

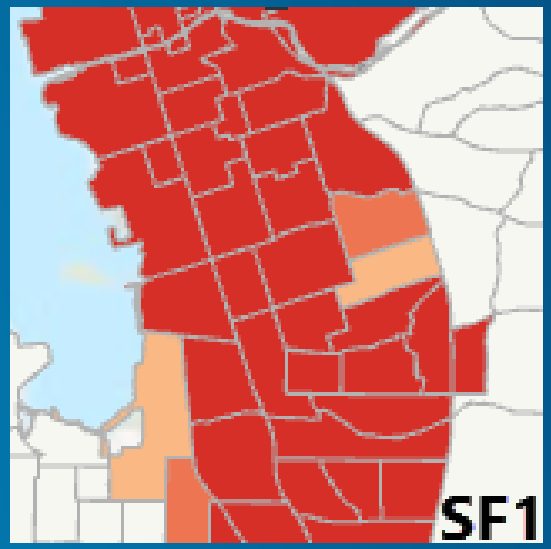


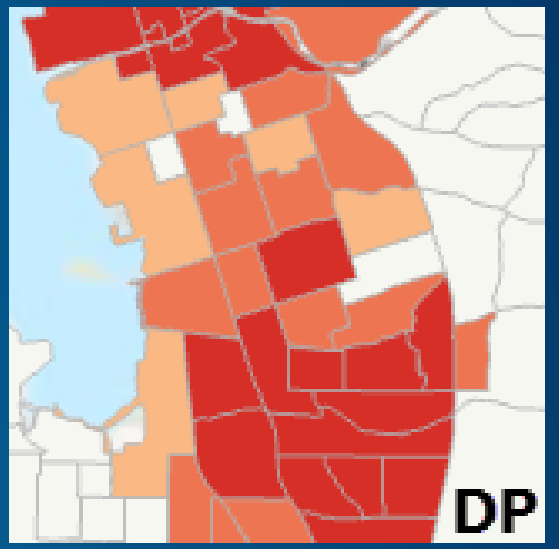
# Mapping Spatial Patterns

***Question: What is being done to ensure that modifications to the data will not change underlying spatial patterns?***



Hot spot analysis of senior densities at the census tract level of analysis



# Dr. John Abowd

# Differential Privacy: What GIS Users Need To Know

**John M. Abowd**

**U.S. Census Bureau**

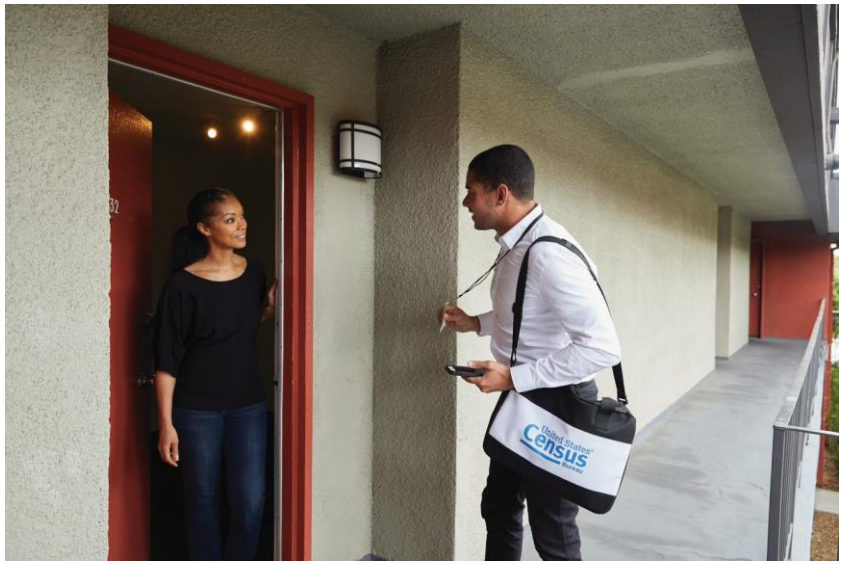
**Esri Webinar July 22, 2020**

The views expressed in this presentation are those of the speaker not the U.S. Census Bureau.

# Our Commitment to Data Stewardship

Data stewardship is central to the Census Bureau's mission to produce high-quality statistics about the people and economy of the United States.

Our commitment to protect the privacy of our respondents and the confidentiality of their data is both a legal obligation and a core component of our institutional culture.



# The Census Bureau's Decision

Advances in computing power and the availability of external data sources make database reconstruction and re-identification increasingly likely.

The Census Bureau recognized that its traditional disclosure avoidance methods are increasingly insufficient to counter these risks.

To meet its continuing obligations to safeguard respondent information, the Census Bureau has committed to modernizing its approach to privacy protections.

# Privacy protection out of the shadows

- Certain privacy practices for previous censuses depended upon obfuscation
- DAS demonstration data are the most transparent view into Census Bureau privacy practices ever
- We appreciate and are excited to assess feedback from our external partners

Start | Base Map | Selection | Results

Save | Load | Feedback | Previous Extent | Hide Tabs | Hide Chart/Report

# Distance/Direction Analysis

*Work to Home*

## Display Settings

Labor Market Segment Filter: All Workers

Year: 2017

## Map Controls

Color Key

Thermal Overlay

Point Overlay

Selection Outline

Identify | Zoom to Selection

Clear Overlays | Animate Overlays

## Report/Map Outputs

Detailed Report

Export Geography

Print Chart/Map

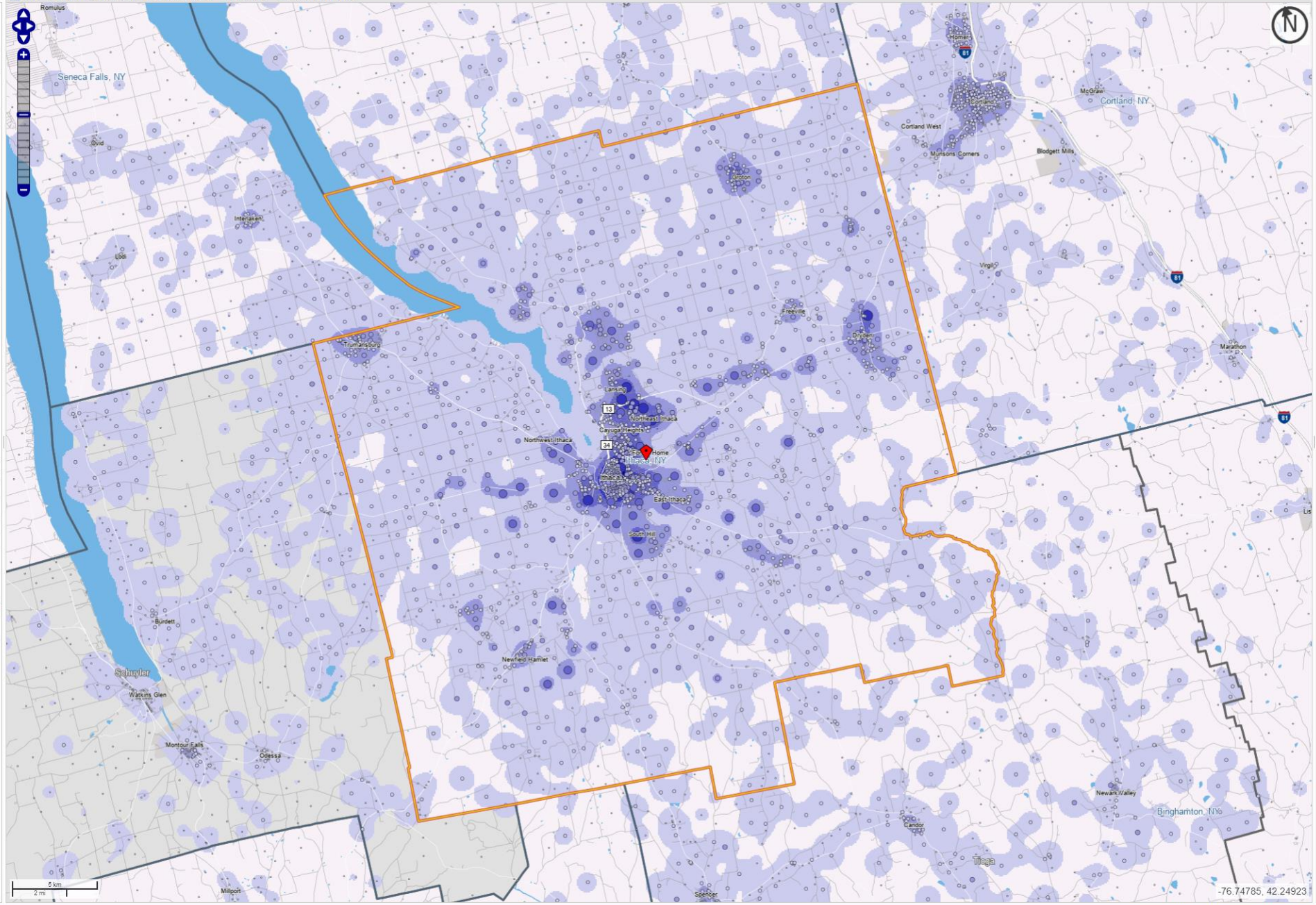
## Legends

- 5 - 99 Jobs/Sq.Mile
- 100 - 382 Jobs/Sq.Mile
- 383 - 854 Jobs/Sq.Mile
- 855 - 1,514 Jobs/Sq.Mile
- 1,515 - 2,364 Jobs/Sq.Mile
- 1 - 3 Jobs
- 4 - 18 Jobs
- 19 - 59 Jobs
- 60 - 139 Jobs
- 140 - 271 Jobs
- Analysis Selection

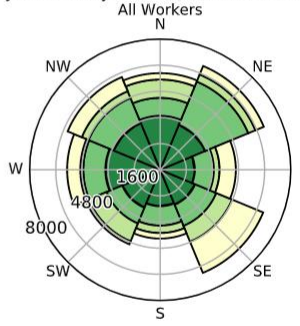
## Analysis Settings

| | |
|---|---|
| Analysis Type | Distance/Direction |
| Selection area as | Work |
| Year(s) | 2017 |
| Job Type | Primary Jobs |
| Selection Area | Ithaca, NY from Metropolitan/Micropolitan Areas (CBSA) |
| Selected Census Blocks | 3,082 |
| Analysis Generation Date | 07/20/2020 16:43 - OnTheMap 6.6 |
| Code Revision | d7f8a300c9f4e458f61bc73d3099ca2cb8f8feaa |
| LODES Data Version | 20170818 |

Change Settings




Job Counts by Distance/Direction in 2017
All Workers




View as: Radar Chart

**Jobs by Distance - Work Census Block to Home Census Block**

| | 2017 Count | Share |
|---|---|---|
| Total Primary Jobs | 45,225 | 100.0% |
| Less than 10 miles | 21,905 | 48.4% |
| 10 to 24 miles | 11,790 | 26.1% |
| 25 to 50 miles | 5,255 | 11.6% |
| Greater than 50 miles | 6,275 | 13.9% |

# Census TopDown Algorithm (TDA): Requirements and Properties I

TDA is the primary formally private 2020 Census disclosure limitation algorithm under development

**Inputs:**

- Post-edits-and-imputation microdata records (Census Edited File – CEF)
- Required structural zeros and data-dependent invariants

**Processing:**

- Convert CEF to an equivalent histogram
- Apply DP measurements and perform mathematical optimizations
- Create noisy histogram; convert back to microdata

**Output:**

Return the Microdata Detail File (the MDF; microdata with same schema as CEF)

**Example:**

- Schema: Geography × Ethnicity × Race × Age × Sex × HHGQ
- This product yields a “histogram” (fully saturated contingency table)
- With shape: ≈ 8M × 2 × 63 × 116 × 2 × 43 = ≈ 8M × 1.25M

# Census TDA: Requirements and Properties II

**Data-dependent invariants:**

Properties of true data that must hold exactly (*no noise*)

**Current data-dependent invariants:**

- State population totals
- Count of occupied GQ facilities by type by block (not population)
- Total count of housing units by block (not population)

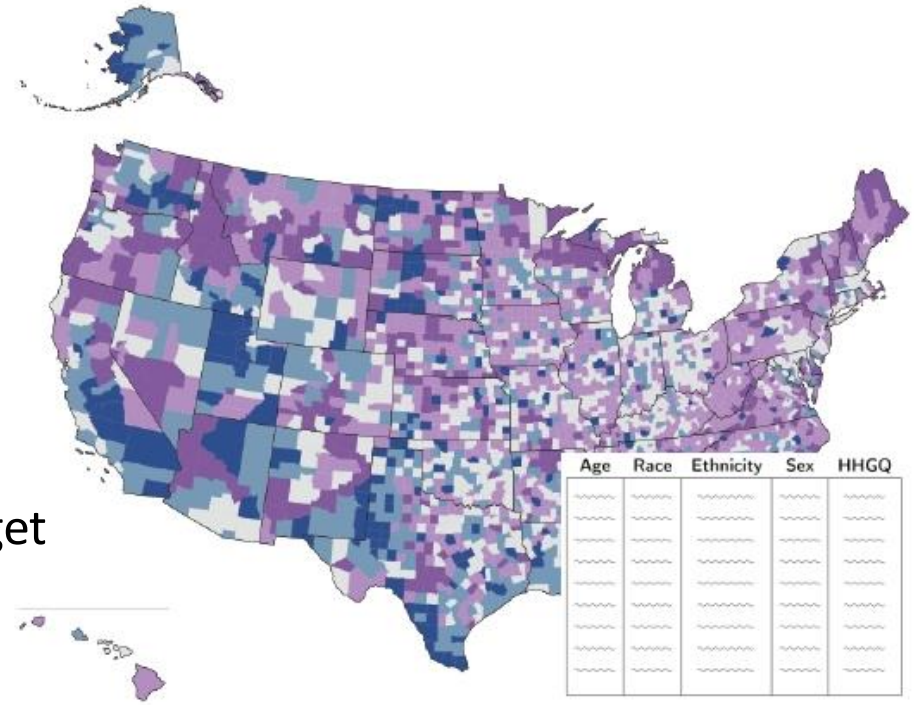
**Utility/Accuracy for pre-specified tabulations**

- Full privacy + full accuracy for arbitrary uses = impossible
- P.L. 94-171: tabulations used for redistricting
- Demographic and Housing Characteristics File
  - Principal successor to 2010 Summary File 1
  - TDA creates separate Person and Housing Unit microdata sets

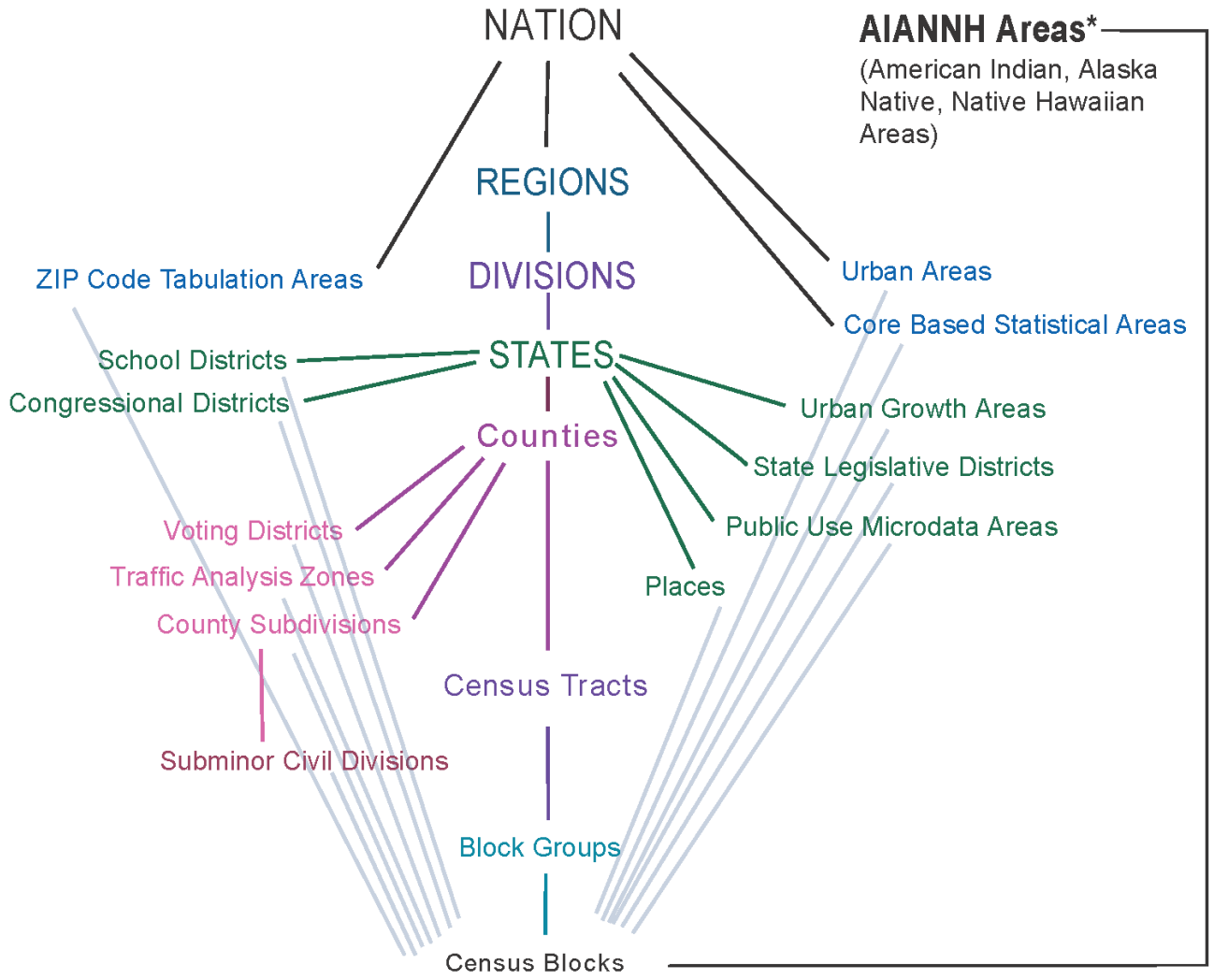
**$\epsilon$-consistency:** error $\rightarrow 0$ as privacy loss $\epsilon \rightarrow \infty$

**Transparency:** source code and parameters made public

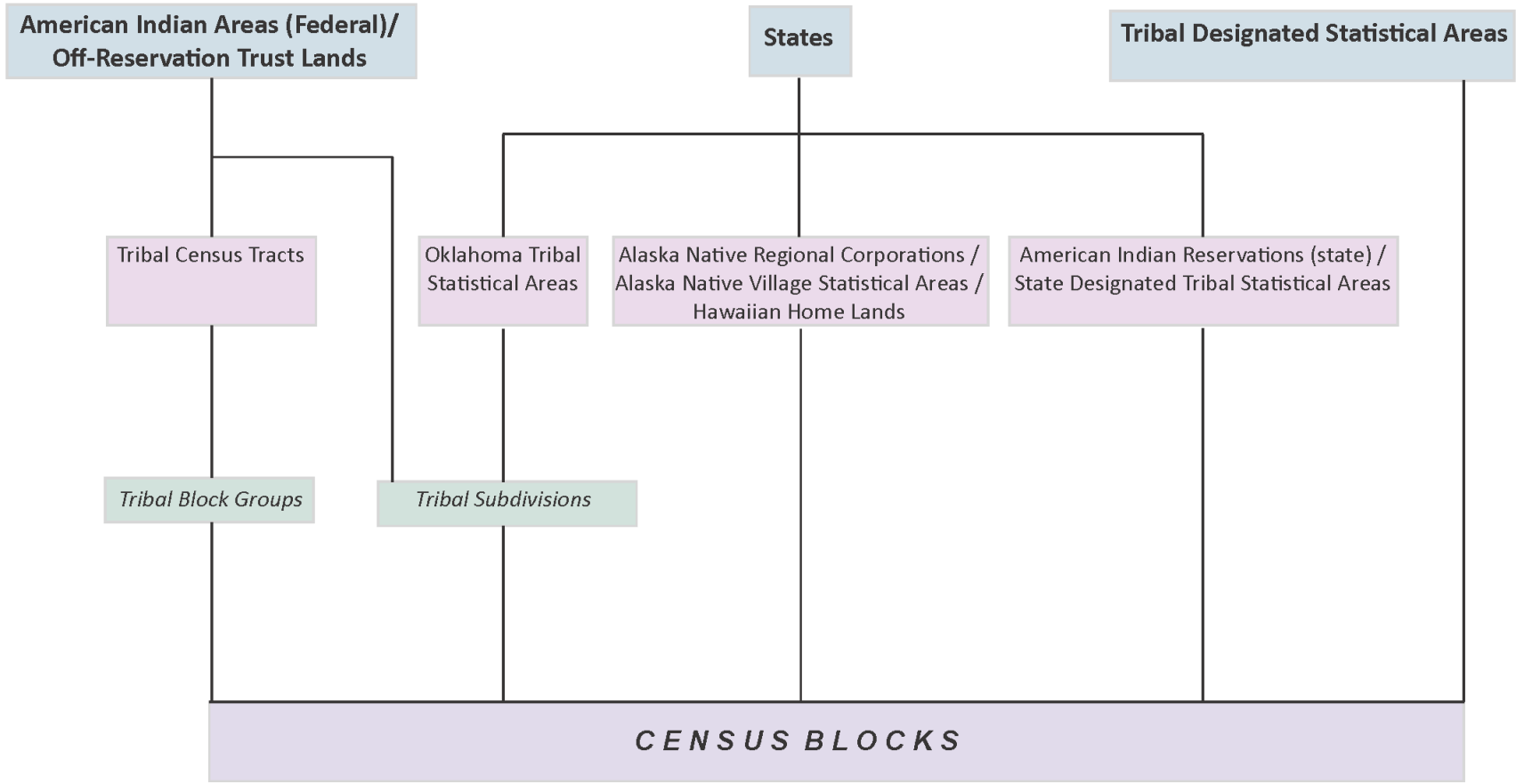
# Basic Structure of TDA




1. Split privacy-loss budget $\varepsilon$ into pieces: $\varepsilon_{nat}, \varepsilon_{state}, \ldots$
2. Ignore geography, make national histogram $\tilde{H}^0$ using $\varepsilon_{nat}$ budget
3. Using $\varepsilon_{state}$ budget, make state histograms: $\tilde{H}^1_{AK}, \tilde{H}^1_{AL}, \ldots, \tilde{H}^1_{WY}$
   - Must be consistent
   - i.e., $\sum_{s \in states} \tilde{H}^1_s = \tilde{H}^0$
4. Recurse down the hierarchy
5. Invariants imposed as constraints in each optimization problem (with notable complications!)

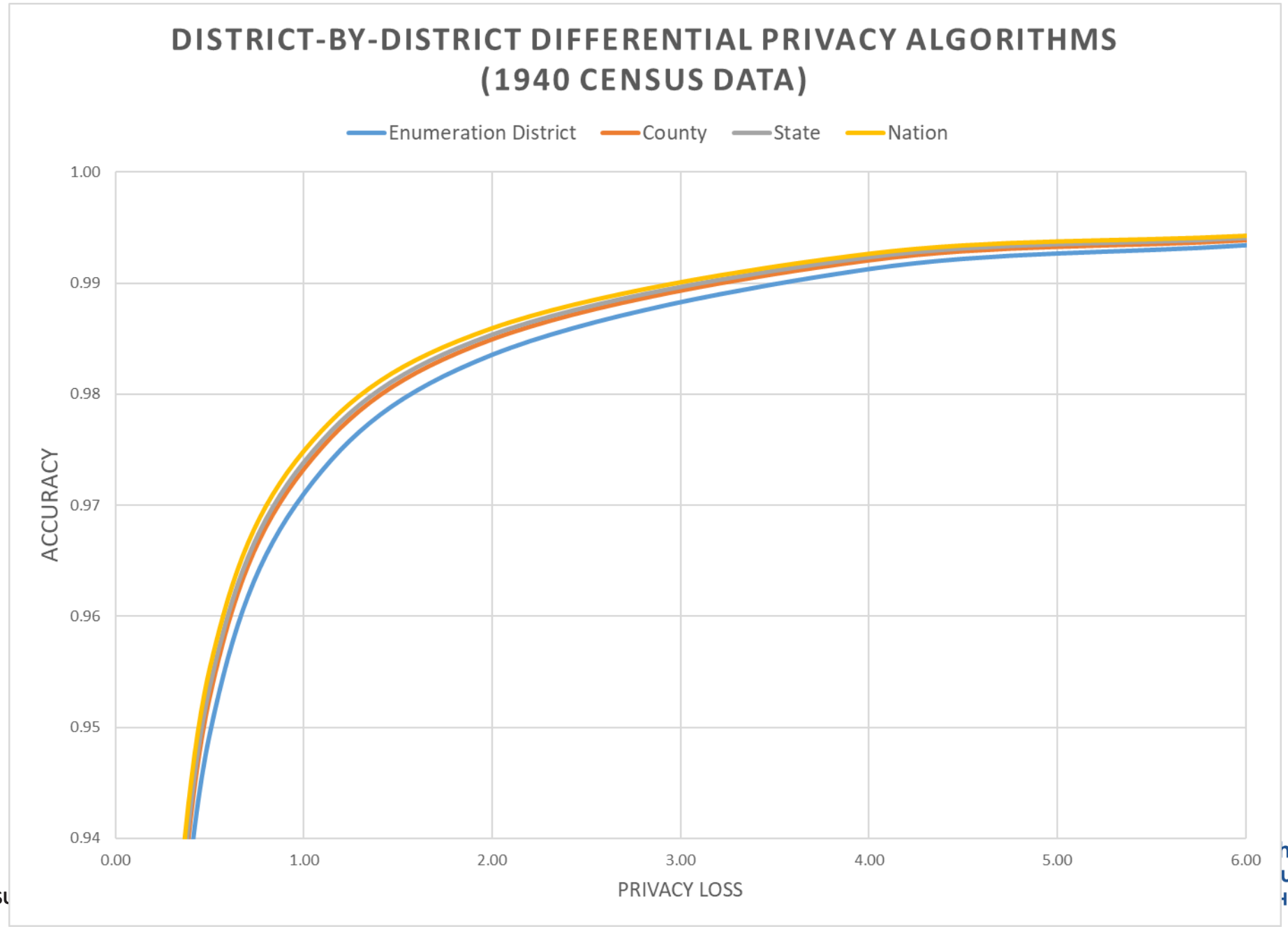


## *Hierarchy of American Indian, Alaska Native, and Native Hawaiian Areas*



# Accurate, but to whom?

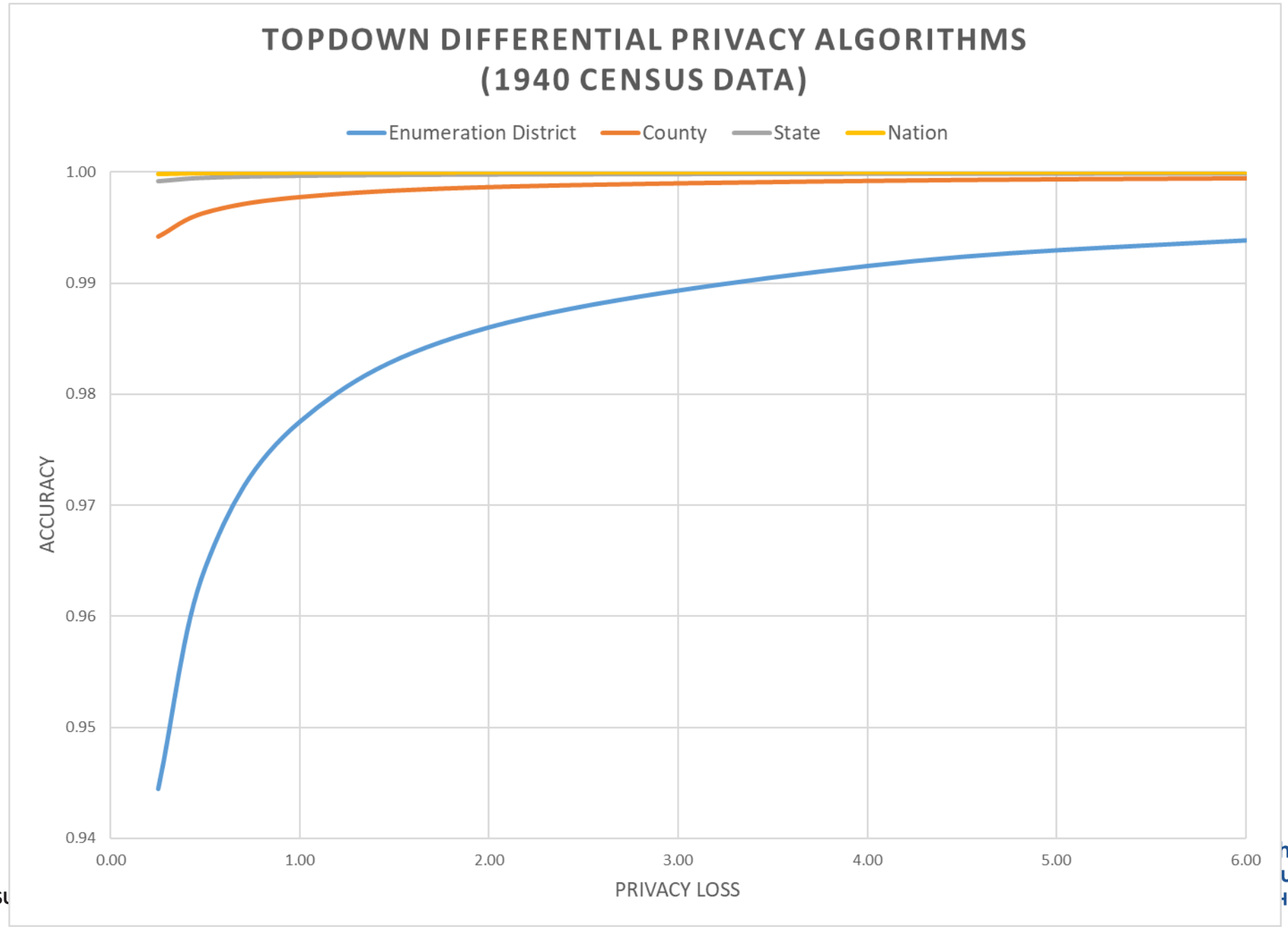
- DAS operates under interpretable formal privacy guarantees, given privacy-loss budgets
- Accuracy properties depend upon the output metric (use case)
- Distinct groups of data users will have a particular analyses they wish to be accurate
- Tuning accuracy for a given analysis can reduce accuracy for other analyses
- Policy makers must consider reasonable overall accuracy metrics for privacy tradeoffs
- Knowing how overall metrics correspond to user results could help optimize DAS






nape
uture
HERE >

United States® Census 2020



# Selected Resources

Technical: https://github.com/uscensusbureau/census2020-das-e2e/blob/master/doc/20190711_0945_Consistency_for_Large_Scale_Differentially_Private_Histograms.pdf

Basics: https://www.census.gov/about/policies/privacy/statistical_safeguards/disclosure-avoidance-2020-census.html

Updates: https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/2020-census-data-products/2020-das-updates.html

**Thank you**

John.Maron.Abowd@census.gov

# Dr. Lauren Scott Griffin

# Question 1 – Accuracy:

- ***Do you have any recommendations to help users determine if the 2020 data they're using (at any geographic scale) is accurate enough for their application?***

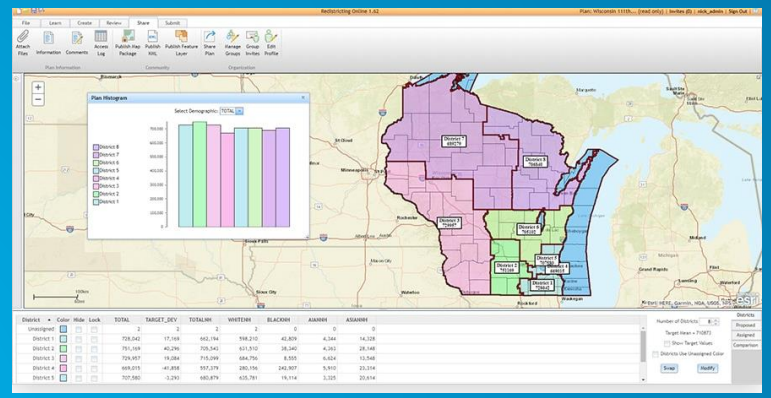
- **Example: We want to know how many kindergarten teachers to recruit next year. The census provides the number of 4-year old's. Will you provide something like MOE to allow analysis of worst/best case scenario, for example?**

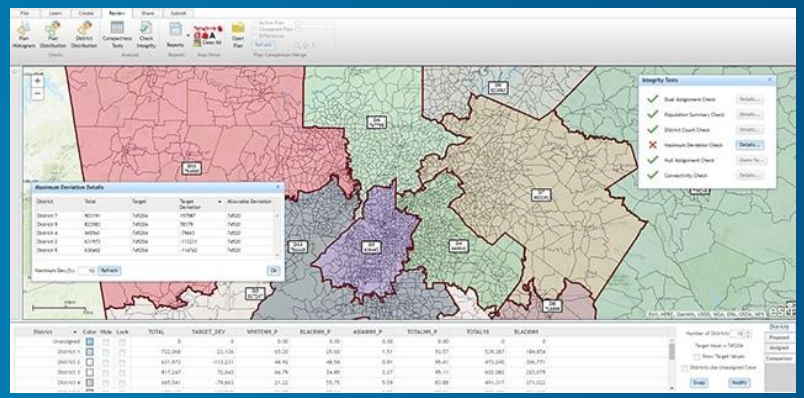
*Photo by Bill Oxford on Unsplash*

# Question 2 – Redistricting:

- ***Will the Census Bureau be able to defend their published counts in court? Do you expect more challenges to sub-state census counts because of the shift to Differential Privacy?***

- **Example: For redistricting, while housing units will be invariant from the counted values, noise will be introduced to population counts used in re-districting.**

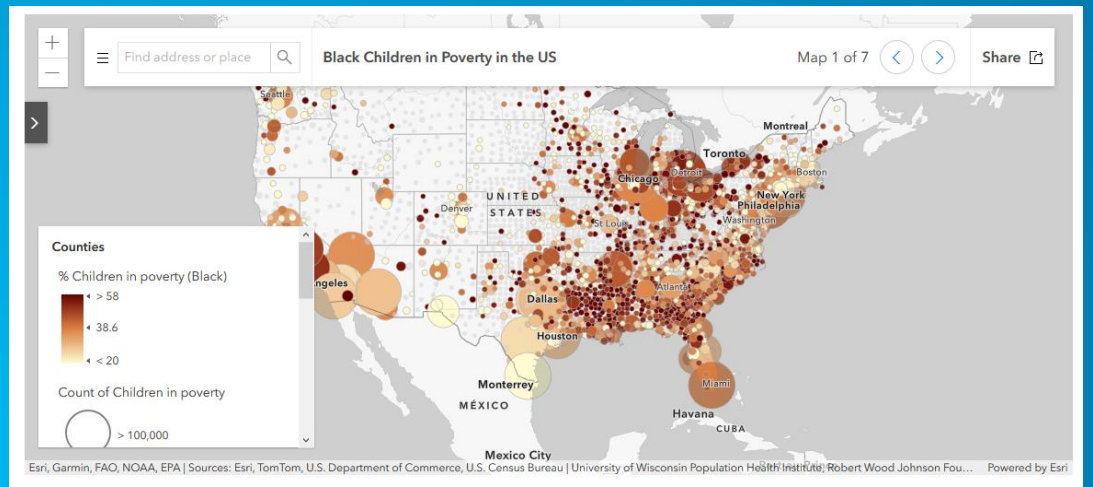


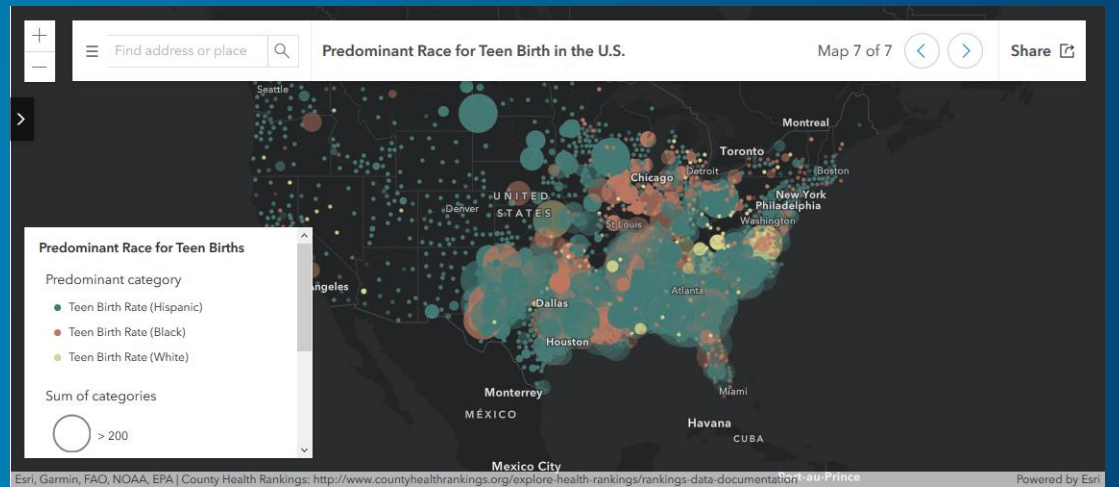


# Question 3 – Marginalized Populations:

- ***The data for people associated with small geographies or small numbers (such as immigrants, Tribal populations, and very specific racial groups) will be less accurate than data for larger groups/geographies. This has implications for social justice and social equity. Is it possible for differential privacy to accurately count these populations while still maintaining individual anonymity?***

- **Example: We need to identify the best places to solicit marrow donors for a rare race/ethnicity. The lives of children depend on finding willing donors quickly and effectively.  The privacy risk and accuracy seem to be at direct odds here.**

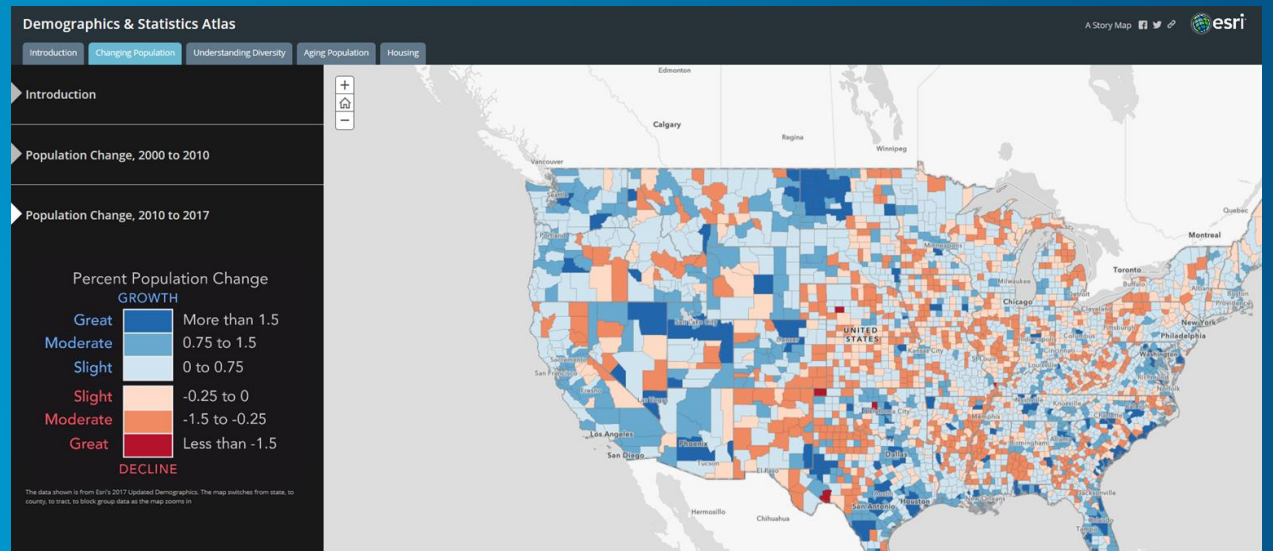


# Question 4 – Temporal Trends:

- ***GIS analysts are often interested in understanding changes over space and time. With the Census Bureau's Disclosure Avoidance System (DAS), will they still be able to evaluate trends for small areas such as block groups, school districts, and voting districts? In other words, what are the implications of the DAS for longitudinal studies?***

- **Example: I'm a store owner and want to evaluate how demographic characteristics around the store have changed over the past 10 year. Can I compare the 2020 data to 2010 data?**



# Question & Answer

Please Enter Questions in the Questions Window

# Resources

- US Census Bureau Disclosure Avoidance and the 2020 Census
- Esri Blog on Census 2020 Differential Privacy
- Developing the DAS: Progress Metrics and Data Runs
- IPUMS 2010 Demonstration Data
- Esri ACS Methodology White Paper
- Esri Demographic User Resources

# Closing Challenge

**What are we willing to give up in order to gain accuracy?**

**What can the geographic community do to provide tools and methods to work with these noisy data to understand patterns?**

**Share your questions and ideas with Esri and US Census teams**

**Privacy – v – Accuracy**